Biomedical Optics EXPRESS

Quantitative characterization of bovine serum albumin thin-films using terahertz spectroscopy and machine learning methods

YIWEN SUN,[1] PENGJU DU,[1] XINGXING LU,[1] PENGFEI XIE,[1] ZHENGFANG QIAN,[2] SHUTING FAN,[2,3,5] AND ZEXUAN ZHU[4,6]

[1]*National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong, Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Department of Biomedical, Engineering, School of Medicine, Shenzhen University, Shenzhen 518060, China*
[2]*College of Electronic Science and Technology, Shenzhen University, Shenzhen 518060, China*
[3]*M013, School of Physics and Astrophysics, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[4]*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China*
[5]*fanshuting1024@gmail.com*
[6]*zhuzx@szu.edu.cn*

Abstract: The development of new spectral analysis methods in bio thin-film detection has generated intense interest in terahertz (THz) spectroscopy and its application in a wide range of fields. In this paper, it is the first time that machine learning methods are applied to the quantitative characterization of bovine serum albumin (BSA) deposited thin-films detected by terahertz time-domain spectroscopy. The spectra data of BSA thin-films prepared by solutions with concentrations ranging from 0.5 to 35 mg/ml are analyzed using the support vector regression method to learn the underlying model of the frequency against the target concentration. The learned mode successfully predicts the concentrations of the unknown test samples with a coefficient of determination $R^2 = 0.97932$. Furthermore, aiming to identify the relevance of each frequency to the concentration, the maximal information coefficient statistical analysis is used and the three most discriminating frequencies in THz frequency are identified at 1.2, 1.1 and 0.5 THz respectively, which means a good prediction for BSA concentration can be achieved by using the top three relevant frequencies. Moreover, the top discriminating frequencies are in good agreement with the frequencies predicted by a long-wavelength elastic vibration model for BSA protein.

OCIS codes: (070.4790) Spectrum analysis; (300.6495) Spectroscopy, terahertz; (170.1420) Biology.

References and links

1. M. Tonouchi, "Cutting-edge terahertz technology," Nat. Photonics **1**(2), 97–105 (2007).
2. S. L. Dexheimer, *Terahertz Spectroscopy: Principles and Applications* (CRC Press, 2007).
3. T. M. Korter and D. F. Plusquellic, "Continuous-wave terahertz spectroscopy of biotin: vibrational anharmonicity in the far-infrared," Chem. Phys. Lett. **385**(1–2), 45–51 (2004).
4. M. R. Kutteruf, C. M. Brown, L. K. Iwaki, M. B. Campbell, T. M. Korter, and E. J. Heilweil, "Terahertz spectroscopy of short-chain polypeptides," Chem. Phys. Lett. **375**(3–4), 337–343 (2003).
5. Y. Sun, Y. Zhang, and E. Pickwell-Macpherson, "Investigating antibody interactions with a polar liquid using terahertz pulsed spectroscopy," Biophys. J. **100**(1), 225–231 (2011).
6. Y. Sun, J. Zhong, C. Zhang, J. Zuo, and E. Pickwell-MacPherson, "Label-free detection and characterization of the binding of hemagglutinin protein and broadly neutralizing monoclonal antibodies using terahertz spectroscopy," J. Biomed. Opt. **20**(3), 037006 (2015).
7. Y. Sun, Z. Zhu, S. Chen, J. Balakrishnan, D. Abbott, A. T. Ahuja, and E. Pickwell-Macpherson, "Observing the temperature dependent transition of the GP2 peptide using terahertz spectroscopy," PLoS One **7**(11), e50306 (2012).
8. J. F. O'Hara, W. Withayachumnankul, and I. Al-Naib, "A review on thin-film sensing with terahertz waves," J. Infrared Millim. Te. **33**(3), 245–291 (2012).
9. F. Severcan and P. Haris, *Vibrational Spectroscopy in Diagnosis and Screening* (IOS Press, 2012).

#313259 https://doi.org/10.1364/BOE.9.002917
Journal © 2018 Received 20 Nov 2017; revised 2 May 2018; accepted 27 May 2018; published 6 Jun 2018

10. M. Hofmann, M. Winzer, C. Weber, and H. Gieseler, “Prediction of protein aggregation in high concentration protein solutions utilizing protein-protein interactions determined by low volume static light scattering,” J. Pharm. Sci. **105**(6), 1819–1828 (2016).
11. W. Withayachumnankul, J. F. O’Hara, W. Cao, I. Al-Naib, and W. Zhang, “Limitation in thin-film sensing with transmission-mode terahertz time-domain spectroscopy,” Opt. Express **22**(1), 972–986 (2014).
12. M. Gocic, D. Petkovic, S. Shamshirband, and A. Kamsin, “Comparative analysis of reference evapotranspiration equations modelling by extreme learning machine,” Comput. Electron. Age. **127**, 56–63 (2016).
13. H. L. Zhan, K. Zhao, H. Zhao, Q. Li, S. M. Zhu, and L. Z. Xiao, “The spectral analysis of fuel oils using terahertz radiation and chemometric methods,” J. Phys. D Appl. Phys. **49**(39), 395101 (2016).
14. H. Zhan, Q. Li, K. Zhao, L. Zhang, Z. Zhang, C. Zhang, and L. Xiao, “Evaluating PM2.5 at a construction site using terahertz radiation,” IEEE T. THZ Sci. Techn. **5**(6), 1028–1034 (2015).
15. J. El Haddad, F. de Miollis, J. Bou Sleiman, L. Canioni, P. Mounaix, and B. Bousquet, “Chemometrics applied to quantitative analysis of ternary mixtures by terahertz spectroscopy,” Anal. Chem. **86**(10), 4927–4933 (2014).
16. J. I. Boye, I. Alli, and A. A. Ismail, “Interactions involved in the gelation of bovine serum albumin,” J. Agric. Food Chem. **44**(4), 996–1004 (1996).
17. J. D. Ferry, “Protein gels,” Adv. Protein Chem. **4**, 1–78 (1948).
18. C. Mircean, I. Shmulevich, D. Cogdell, W. Choi, Y. Jia, I. Tabus, S. R. Hamilton, and W. Zhang, “Robust estimation of protein expression ratios with lysate microarray technology,” Bioinformatics **21**(9), 1935–1942 (2005).
19. H. J. Willison, K. Townson, J. Veitch, J. Boffey, N. Isaacs, S. M. Andersen, P. Zhang, C. C. Ling, and D. R. Bundle, “Synthetic disialylgalactose immunoadsorbents deplete anti-GQ1b antibodies from autoimmune neuropathy sera,” Brain **127**(3), 680–691 (2003).
20. L. K. Gifford, L. G. Carter, M. J. Gabanyi, H. M. Berman, and P. D. Adams, “The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: A Structural Biology Web Resource,” J. Struct. Funct. Genomics **13**(2), 57–62 (2012).
21. Y. Y. Studentsov, M. Schiffman, H. D. Strickler, G. Y. Ho, Y. Y. Pang, J. Schiller, R. Herrero, and R. D. Burk, “Enhanced enzyme-linked immunosorbent assay for detection of antibodies to virus-like particles of human papillomavirus,” J. Clin. Microbiol. **40**(5), 1755–1760 (2002).
22. H. Yoneyama, M. Yamashita, S. Kasai, K. Kawase, R. Ueno, H. Ito, and T. Ouchi, “Terahertz spectroscopy of native-conformation and thermally denatured bovine serum albumin (BSA),” Phys. Med. Biol. **53**(13), 3543–3549 (2008).
23. B. S. Kalanoor, M. Ronen, Z. Oren, D. Gerber, and Y. R. Tischler, “New Method to Study the Vibrational Modes of Biomolecules in the Terahertz Range Based on a Single-Stage Raman Spectrometer,” ACS Omega **2**(3), 1232–1240 (2017).
24. A. G. Markelz, A. Roitberg, and E. J. Heilweil, “Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz,” Chem. Phys. Lett. **320**(1-2), 42–48 (2000).
25. J. Xu, K. W. Plaxco, and S. J. Allen, “Probing the collective vibrational dynamics of a protein in liquid water by terahertz absorption spectroscopy,” Protein Sci. **15**(5), 1175–1181 (2006).
26. O. Sushko, R. Dubrovka, and R. S. Donnan, “Sub-terahertz spectroscopy reveals that proteins influence the properties of water at greater distances than previously detected,” J. Chem. Phys. **142**(5), 055101 (2015).
27. M. Mernea, A. Leca, T. Dascalu, and M. Dan, *Bovine Serum Albumin 3D Structure Determination by THz Spectroscopy and Molecular Modeling* (Springer, 2011), pp. 101–105.
28. P. H. Bolívar and R. Sczech, “THz spectroscopy of bovine serum albumin solution using the long-range guided mode supported by thin liquid films,” in *CLEO: 2014*, OSA Technical Digest (online) (Optical Society of America, 2014), paper SF1F.4.
29. J. Song, L. C. Chen, and B. J. Li, “Super-sensitive optical biosensor with a spectrometer on a chip,” IEEE Biotechnology & Biotechnological Equipment **27**(4), 4040–4043 (2017).
30. S. Krimi, G. Torosyan, and R. Beigang, “Advanced GPU-Based Terahertz Approach for In-Line Multilayer Thickness Measurements,” IEEE J. Sel. Top. Quant. **23**(4), 8501112 (2017).
31. T. Yasuda, T. Iwata, T. Araki, and T. Yasui, “Improvement of minimum paint film thickness for THz paint meters by multiple-regression analysis,” Appl. Opt. **46**(30), 7518–7526 (2007).
32. T. Iwata, S. Yoshioka, S. Nakamura, Y. Mizutani, and T. Yasui, “Prediction of the Thickness of a Thin Paint Film by Applying a Modified Partial-Least-Squares-1 Method to Data Obtained in Terahertz Reflectometry,” J. Infrared Milli. Terahz. Waves **34**(10), 646–659 (2013).
33. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
34. A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, “Support vector regression with chaos-based firefly algorithm for stock market price forecasting,” Appl. Soft Comput. **13**(2), 947–958 (2013).
35. Y. Ren, P. N. Suganthan, and N. Srikanth, “A novel empirical mode decomposition with support vector regression for wind speed forecasting,” IEEE Trans. Neural Netw. Learn. Syst. **27**(8), 1793–1798 (2016).
36. L. Q. Hou, S. L. Yang, and Z. Q. Chen, “The use of data mining techniques and support vector regression for financial forecasting,” Int. J. Database Theory Appl. **6**(4), 145–156 (2013).
37. A. Sanchez-Gonzalez, P. Micaelli, C. Olivier, T. R. Barillot, M. Ilchen, A. A. Lutman, A. Marinelli, T. Maxwell, A. Achner, M. Agåker, N. Berrah, C. Bostedt, J. D. Bozek, J. Buck, P. H. Bucksbaum, S. C. Montero, B. Cooper, J. P. Cryan, M. Dong, R. Feifel, L. J. Frasinski, H. Fukuzawa, A. Galler, G. Hartmann, N. Hartmann, W. Helml, A. S. Johnson, A. Knie, A. O. Lindahl, J. Liu, K. Motomura, M. Mucke, C. O’Grady, J. E. Rubensson, E. R.

Simpson, R. J. Squibb, C. Såthe, K. Ueda, M. Vacher, D. J. Walke, V. Zhaunerchyk, R. N. Coffee, and J. P. Marangos, "Accurate prediction of X-ray pulse properties from a free-electron laser using machine learning," Nat. Commun. **8**, 15461 (2017).

38. S. Ubaru, A. Międlar, Y. Saad, and J. R. Chelikowsky, "Formation enthalpies for transition metal alloys using machine learning," Phys. Rev. B **95**(21), 214102 (2017).
39. A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," Stat. Comput. **14**(3), 199–222 (2004).
40. C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," ACM T. Intel. Syst. Tec. **2**(3), 27 (2011).
41. D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," Science **334**(6062), 1518–1524 (2011).
42. S. Ihara, *Information Theory for Continuous Systems* (World Scientific, 1993).
43. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, 2009).
44. S. I. Bastrukov, "Low-frequency elastic response of a spherical particle," Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **49**(4), 3166–3170 (1994).
45. S. Perticaroli, J. D. Nickels, G. Ehlers, H. O'Neill, Q. Zhang, and A. P. Sokolov, "Secondary structure and rigidity in model proteins," Soft Matter **9**(40), 9548–9556 (2013).
46. E. Mylonas and D. I. Svergun, "Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering," J. Appl. Cryst. **40**(s1), s245–s249 (2007).

1. Introduction

Over the past decade, the simple, label free, and high sensitive protein detection techniques have been extensively investigated. Terahertz time-domain spectroscopy (THz-TDS) is one branch of these efforts that has been advancing rapidly in recent years. Because of the low photon-energy [1], high signal-to-noise ratio [2] and molecule resonance responses [3,4], the terahertz spectrum hosts a range of important microscopic phenomena of biomolecular interactions [5–7]. Thin films detection which is specialized to enable successful sensing for a small amount of sample (e.g. protein, DNA) has potential benefits to broaden THz-TDS bio-applications [8]. Quantitative analysis has generated intense interest in a wide range of fields, including protein structure prediction and formulations optimization [9], cell culture conditions controlling and monitoring by improving target protein production [10]. However, quantitative characterization of the bio thin-films in terahertz frequency has not been intensely studied, since the interaction length between terahertz waves and a sample film is short that the extracted optical parameters are not reliable [11]. The demand for new terahertz spectra analysis methods in bio thin-film detection has increased significantly. Machine learning methods are capable of learning the underlying model of the experimental data and generalizing well to unknown test data. Therefore, they suit the requirements of data analysis for laboratory and industry purpose [12–15]. In this study, a machine learning framework is proposed to successfully predict the function of the frequencies and the target concentrations for an exemplar protein (bovine serum albumin protein) thin-film detected by the terahertz spectroscopy.

Bovine serum albumin (BSA) is a serum albumin protein containing 583 amino acid residues with a molecular weight of 66.430 kDa [16,17]. It is a multifunctional and low-cost protein which is able to block the nonspecific binding sites during protein-protein interactions. Therefore, BSA has been widely used in various biochemical detection techniques such as ELISAs (enzyme-linked immunosorbent assay) [18,19], immunohistochemistry [20] and immunoblots [21]. Moreover, by comparing an unknown quantity of protein to known amounts of BSA, it is often used as a protein concentration standard, which is therefore of great importance to identify BSA quantitatively and qualitatively. Even though BSA in a solid state (pressed pellet) and in solution has been previously investigated using THz spectroscopy [22–28], a great challenge for the property identification of BSA thin-films is obvious because subwavelength sample thicknesses impose great difficulties to conventional terahertz spectroscopy [29]. This work presents the first THz time-domain spectroscopy investigation of the BSA thin-films with a support vector regression (SVR) method to learn the function of the frequency and the target concentration. Comparing most of the previous prediction methods which are based on the thin-film thickness measurements [30–32], accurate quantitative prediction of unknown samples can be achieved by the learned function in the SVR model, without the film thickness discussion. Furthermore, SVR model applied to THz data in this work allows one to take into account possible nonlinearities in the detected signals to identify protein concentrations. Finally, the maximal information coefficient (MIC) was applied to identify the most discriminating frequencies to concentrations of BAS in THz region, which correspond to the fundamental vibration frequencies of a long-wavelength elastic vibration model.

2. Experimental method

2.1 Sample preparation

Double-side-polished 0.5 mm thick quartz substrates were ultrasonically cleaned for 10 min successively in acetone, isopropyl alcohol, and deionized (DI) water and then surface treated by O_2 plasma for 5 min to improve the hydrophilicity. Prior to BSA deposited thin-films preparation, a BSA stock solution was made by dissolving 3.5 g of solid BSA powder (A3912; Sigma-Aldrich, St. Louis, MO) into 100 ml of DI water. The stock solution was further diluted to obtain 21 different concentrations (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 12.0, 14.0, 16.0, 18.0, 20.0, 25.0, 30.0, 35.0 mg/ml). The aqueous solutions at room temperature were clear and without precipitates. After being stirred for several minutes to improve the uniformity of the BSA thin-films, the BSA solutions at various concentrations were pipetted onto each quartz substrate with same volume by 20 ul. Each sample was prepared in a constant temperature and humidity laboratory with a standard approach so that the error for the actual concentration can be ignored. Liquid thin films were spin coated on quartz substrate by controlling the spin coating speed. To improve the crystallinity, the BSA thin-films were equilibrated for at least 30 minutes in a nitrogen atmosphere. The quantity of deposited BSA protein increased with increasing BSA concentration.

2.2 THz-TDS measurement

THz-TDS measurements were performed using free-space THz-TDS system in the transmission geometry. The system consists of 300 mW in mode-lock operation, 800 nm center wavelength and 84 MHz repetition rate pulse generated by a Ti:sapphire oscillator which is pumped by a 2.2 W 532 nm Nd:YV04 laser (Sprout Lighthouse Photonics). A GaAs semiconductor antenna is used for the THz pulse generation and a ZnTe crystal is employed for electro-optical detection. THz spectra were recorded from 0 to 3.3 mm (equal to a time window ranging from 0 to 22 ps), with a scan speed of 5 µm per step and an interval time of 300 ms, resulting a nominal resolution of 45 GHz. All samples were fabricated on a sample holder with a circular area by ~3 mm in diameter. The optics was purged using nitrogen gas to remove the water vapor from the air to decrease the humidity down to less than 5%. The usable frequency range of the system is from 0.1 to 2.6 THz. Each sample was measured 7 times in order to minimize the random errors produced by the system, as well as present heterogeneities in the sample.

2.3 Machine learning methods

2.3.1 Data denoising with principal component analysis (PCA)

The system uncertainties are very influential for thin-film detection, which results in inevitably noise in the data set. The data processing procedure must be carefully chosen to avoid any deceptive result. A denoising method based on PCA is preformed firstly in this work. The scores of the 7 measurements for each concentration on the first two principal components are obtained. In the two-dimensional space, the centroid coordinates of these 7 measurements are calculated, and the Euclidean distance of each measurement to the centroid is acquired. Then the 7 distance values are tested in T-test with a right-tailed hypothesis at 1%

significance level. We did one-sample Kolmogorov-Smirnov test to test the null hypothesis that the distance data comes from a standard normal distribution at the 5% significance level. The result suggests that the data is normally distributed. We also have performed ANOVA to find that the variances of the distance values in different concentration are significantly different. A measurement with significantly larger distance to the centroid is considered as an outlier or noise data as shown in Fig. 1. Based on PCA-denoising method, 12 outlier points are removed and we finally get a new data set with 135 row measurements, and 43 columns frequencies.

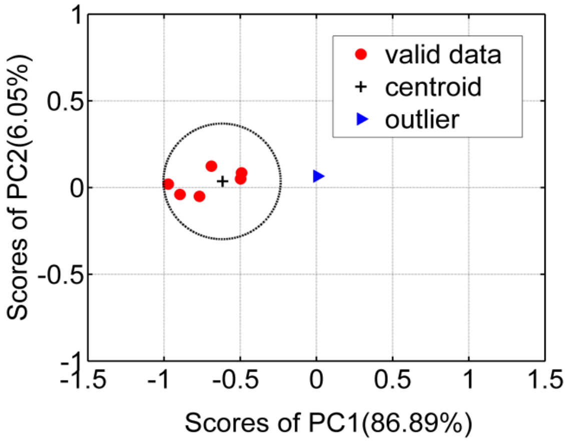

Fig. 1. Schematic diagram of PCA-based denoising method. The scores plot of the first two principal components of the 7 measurements at the concentration of 6 mg/ml.

2.3.2 Spectrum regression analysis by SVR

After PCA-denoising processing, a spectrum regression analysis is performed using SVR algorithm. SVR is a universal regression method inheriting merits from support vector machines [33–35], e.g., the minimization of the structural risk, superiority of generalization for future test data, and ease of handling nonlinear problems with kernel trick. SVR has been successfully used in many fields such as time series prediction [36], X-ray pulse properties prediction [37], and material thermodynamic property prediction [38].

Given a training data set $\{(x_i,y_i)|\ x_i \in R^m,\ y_i \in R^1, i \in\ [1]\}$ of n instances, where each instance (x_i,y_i) consists of an m-dimensional vector $x_i \in R^m$ (m = 43) indicating the values of a measurement in 43 frequencies and a target concentration y_i. The goal of SVR is to find a function $f(x)$ of the frequencies against the concentrations y of BSA, such that all the training instances can be predicted with no more than a predefined deviation $\varepsilon \geq 0$ from the actual targets y and meanwhile $f(x)$ is as flat as possible.

In SVR, a generic form of $f(x)$ is defined as follows:

$$f(x) = w \bullet \Phi(x) + b \tag{2}$$

where w is a weight vector, $b \in$ R is a bias term, $\Phi(x)$ is a mapping function that maps x to a higher-dimensional space if nonlinear regression is considered otherwise $\Phi(x) = x$, and $w \bullet \Phi(x)$ calculates the dot production of w and $\Phi(x)$. The flatness of $f(x)$ can be ensured by minimizing the Euclidean norm $||w||^2$. A prediction on x_i, i.e., $f(x_i)$, is considered accurate if $|f(x_i)\text{-}y_i| \leq \varepsilon$. In practice, to allow deviation violation to some reasonable extent, two slack variables $\xi_i \geq 0$ and $\xi_i{}^* \geq 0$ are usually introduced, such that

$$y_i - f(x_i) \leq \varepsilon + \xi_i \tag{3}$$

$$f(x_i) - y_i \leq \varepsilon + \xi_i^* \tag{4}$$

where regression errors are tolerated up to the value of ξ_i and ξ_i*. The solving of $f(x)$ can be formulated as a convex optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{5}$$

subject to Eq. (3) and (4). The positive constant C controls the trade-off between the flatness of $f(x)$ and the tolerance of the deviation violation. The minimization problem in Eq. (5) can be solved more easily in its dual formulation with kernel trick [33]. In this study, a radial basis function kernel [39] is used. Once $f(x)$ is solved, an unknown test instance can be predicted by inputting the 43 frequency values of this instance to $f(x)$ to get the function output value. For our experiments, the epsilon-SVR model implemented in LIBSVM library [40] is applied and the parameters of the model are selected following ref [40].

2.3.3 Discriminating frequencies identification using MIC

Beyond the regression study using SVR mode, we are also curious about the relevance or discriminability of the 43 frequencies to the target concentrations of BSA in the terahertz region. Thanks to the capability of identifying relationships between two variables and capturing a wide range of variable associations, MIC analysis [41] is taken into account. The basic idea of MIC is to use binning as a means to apply mutual information [42] on random variables. Let variable F denote a frequency of the data and variable Y denotes the concentration, in MIC, the ordered F values and Y values are divided into a bins and b bins, respectively, which results in an a-by-b grid G. The distribution of the values in F-Y space located in the cells of G is denoted as $(F,Y)|_G$. Different grid partitions lead to different distributions. The statistic MIC is the maximum value of the characteristic matrix $M(F,Y)$ defined as follows:

$$M\left(F,Y\right)_{(a,b)} = \frac{\max I\left(\left(F,Y\right)|_G\right)}{\log\min\left\{a,b\right\}} \tag{6}$$

where max$I((F,Y)|_G)$ denotes the maximal mutual information of $(F,Y)|_G$ over all possible grids G. $I(F,Y)|_G$ calculates the mutual information of the probability distribution induced on the cells of a grid G, where the probability of a cell is the proportion of the value falling within the cell. Based on Eq. (6), we can obtain the MIC of F and Y as follows:

$$\mathrm{MIC}\left(F,\mathrm{Y}\right) = \max_{ab<B(n)}\left\{M\left(F,Y\right)_{(a,b)}\right\} \tag{7}$$

where n is the number of instances, $B(n)$ is a function of n, which imposes an upper bounds on the sizes of G for searching the MIC value. In this study, $B(n) = n^{0.4}$ is applied. A larger MIC(F,Y) value indicates that F is more relevant to Y, i.e., F is supposed to be more discriminating in regard to Y.

3. Results and discussion

3.1 THz transmission spectra determination

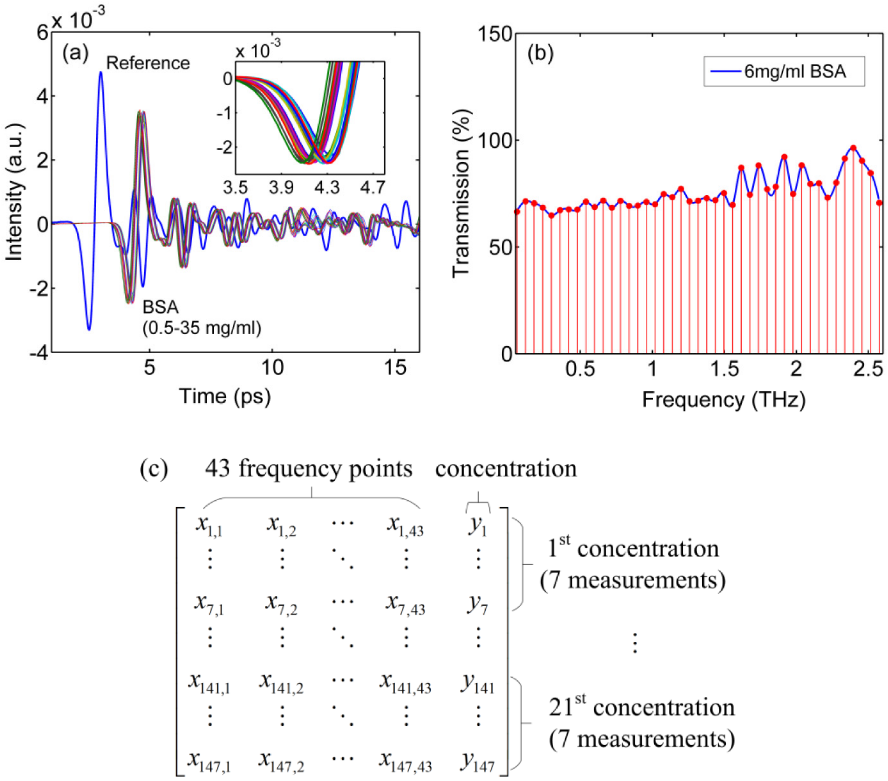

Fig. 2. (a) Time dependence THz intensity spectra of reference and BSA thin-films prepared by the protein aqueous solutions in the concentration range from 0.5 to 35 mg/ml; (b) The transmission spectrum for BSA thin-film prepared by the protein solution of 6 mg/ml from 0.1 to 2.6 THz; (c) illustration of the data matrix with 147 rows (21 concentrations * 7 measurements) and 44 columns (43 frequencies + 1 concentration value).

Terahertz time-domain waveforms were recorded for BSA protein layer on quartz substrate by *Es*(*t*). The clean homogeneous empty quartz surface was measured as a reference *Er*(*t*) by moving the sample holder, enabling the spectroscopic properties of the sample to be accurately determined. By fast Fourier transforming, the frequency spectra of the sample *Es*(ω) and reference *Er*(ω) were obtained. The amplitude of the complex transmission coefficient of the sample *Ts*(ω) can be described by dividing the signal with the sample by the signal without the sample as detailed in Eq. (1).

$$T_s(\omega) = abs\left[\frac{FFT(E_s(t))}{FFT(E_r(t))}\right] \tag{1}$$

Figure 2(a) shows the raw terahertz intensity signal in the time domain of BSA thin-films prepared by the protein aqueous solutions in the concentration range from 0.5 to 35 mg/ml (21 different concentrations). Because the sample thickness is much smaller than the wavelength (1 THz = 300 µm), the deviation of time dependent terahertz spectra is very small. The inset picture zoomed in Fig. 2(a) shows the terahertz intensity peaks which are different from each other, presenting the applicability of THz-TDS can be extended to thin-film sensing. Terahertz transmission spectra of BSA thin-films with various quantities of protein are calculated according to Eq. (1). As an example, the transmission spectrum for

BSA thin-film prepared by a concentration of 6 mg/ml is shown in Fig. 2(b). The time window of about 17 ps was used in this measurement considering the thickness of the sample. Therefore, 43 sampling points (shown as red dots) from 0.1 to 2.6 THz with a consistent frequency step (~58.9 GHz) are selected in order to present the characteristic responses of samples to frequency sufficiently. The semi-periodic oscillations in Fig. 1(b) were caused by the multiple reflection effect at the surface of the quartz substrate (0.5 mm thick) based on our basic calculation. The oscillation can be ignored as the amplitude of the complex transmission coefficient of the sample $Ts(\omega)$ can be defined as the ratio of sample and reference as detailed in Eq. (1). Figure 2(c) demonstrates the way to create the data in matrix format. Making 7 measurements for each sample, 147 transmission spectra each characterized by 43 sampling points and 1 concentration value are preprocessed with PCA for data denoising and then input to SVR model for further investigation. The PCA is performed on the matrix excluding the last y column, so we have more observations than degrees of freedom.

3.2 SVR prediction and performance evaluation

3.2.1 LOOCV scheme

Leave-one-out cross validation (LOOCV) scheme is considered approximately unbiased for estimating the true (expected) prediction errors of machine learning methods [43], therefore, in this study LOOCV is firstly used to evaluate the performance of the SVR model. In LOOCV, each time an instance is selected from the original data set as the test data, and the remaining instances serve as the training data. SVR is trained with the training data and tested on the left out instance to get the deviation. The procedure is repeated until each instance in the data set is tested once and the performance of SVR is averaged over all instances.

The 135 denoised instances are introduced to the LOOCV-SVR model, and the predicted concentrations against the actual values are plotted in Fig. 3. The distributions of actual and predicted concentrations in LOOCV are shown in Fig. 3(a), and the fitting results of the predicted concentrations against the actual values are shown in Fig. 3(b), where a Y = X line is also provided as the reference. Note that the closer the scatter plots are to the reference line, the more reliable is the predictions from the regression model. The error analysis for prediction can also be quantitatively evaluated with the decision coefficient (R^2) and mean square error (MSE). $R^2 \leq 1$ is the correlation coefficient of the predicted values and the actual values. MSE≥0 is the mean square deviation between the predicted values and the actual values. Larger R^2 values (close to one) and smaller MSE values (close to zero) indicate better. As shown in Fig. 3 (a), the LOOCV-SVR model produces a result of $R^2 = 0.97272$ and MSE = 0.015865 on 21 concentration in the range between 0.5 to 35 mg/ml. The inset figure in Fig. 3 (b) presents a clearer vision of fitting results for the lower concentrations (0.5-5 mg/ml). The predicted values estimated using the LOOCV-SVR model are found to be in close agreement with the actual values.

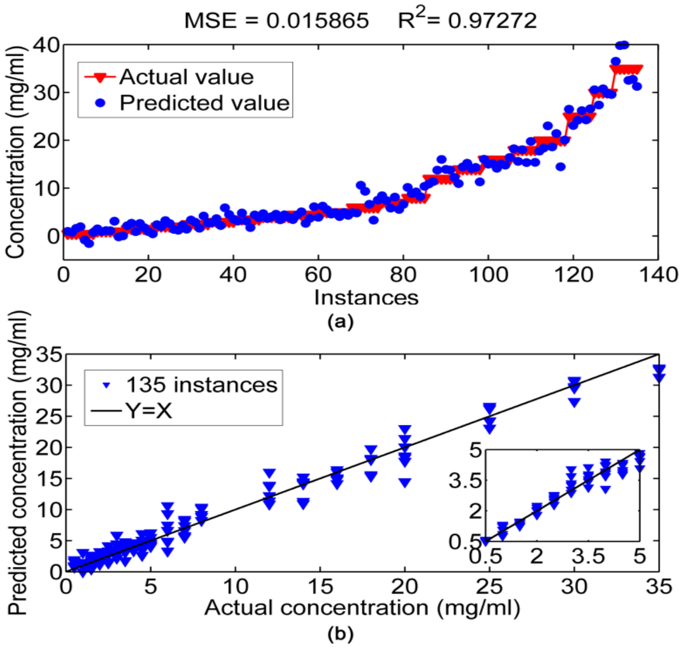

Fig. 3. LOOCV-SVR prediction for various concentrations in the range from 0.5 to 35 mg/ml. (a) The distributions of actual and predicted concentrations in LOOCV, the valid value of x axes is from 1 to 135 represents the 135 spectroscopy measurements; (b) The actual concentrations against the predicted concentrations in LOOCV.

3.2.2 Hold-out validation scheme

In the LOOCV scheme, one test instance is left out in each round to validate the SVR model trained with the remaining instances. Instances of the same concentration with the test one are actually involved in the training set, which could ease the prediction of SVR on the test instance. However, the detected concentration in real-world may not be included in training concentration sequences for prediction.

In order to test the accuracy of SVR model in scenarios where a test instance has no exemplar of the same concentration in the training data, a hold-out validation scheme is subsequently used. Particularly, in each run of the validation, all instances from one concentration are held out as the test data, and the remaining instances from other concentrations serve as the training data. SVR is trained on the training data and tested on the hold-out instances from a totally unknown concentration. The procedure is repeated until each concentration is tested once and the performance of SVR is attained by averaging over all test instances. Since no closer exemplars existing in the training data for a test instance, the prediction performance of SVR in hold-out validation is reasonably believed to be poorer than in LOOCV. The prediction results of SVR using hold-out validation are shown in Fig. 4. It is observed that the prediction in hold-out scheme is less accurate than LOOCV scheme, yet the predicted values fit the actual values with acceptable accuracy (R^2 = 0.91651 and MSE = 0.051639), which suggests the feasibility of the proposed framework for real-world concentration prediction. A classifier like SVR trained with the discrete set of data can work relatively well on a continuum if it correctly captures the underlying distribution of the data. Moreover, the accuracy of the framework can be further improved as long as more measurements for more concentrations are prepared to train the SVR, so that any new instance could find highly similar exemplars in the training data.

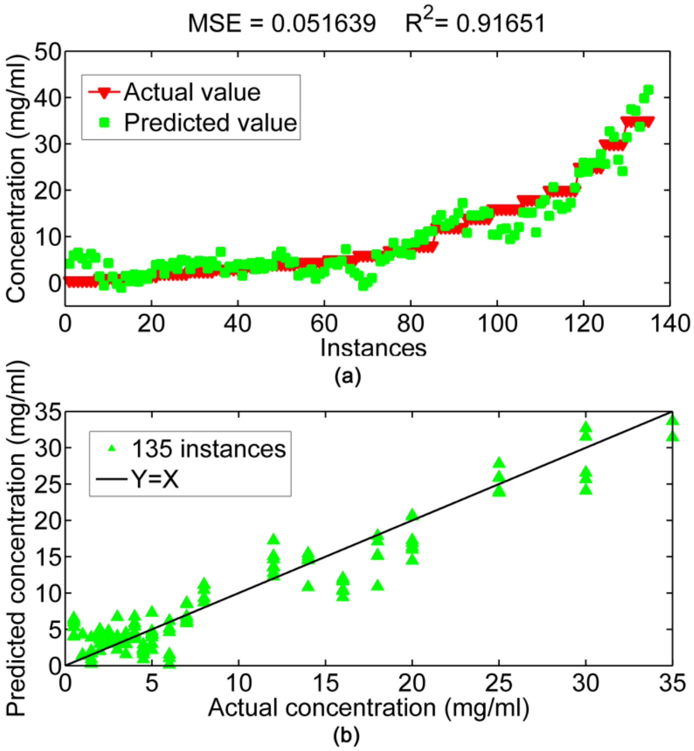

Fig. 4. The prediction results using hold-out validation for various concentrations in the range from 0.5 to 35 mg/ml. (a) The distributions of actual and predicted concentrations in hold-out validation, the valid value of x axes is from 1 to 135 represents the 135 spectroscopy measurements; (b) The actual concentrations against the predicted concentrations in hold-out validation.

3.3 Identification of discriminating frequencies to concentrations

To identify the most discriminating/relevant frequencies related to the concentration, the MIC values of the 43 frequencies against the concentrations are calculated and sorted in descending order. Table 1 presents the top five relevant frequencies based on MIC values. In addition, the top *k* relevant frequencies are selected to test the performance of SVR in LOOCV scheme. As shown in Fig. 5, the R^2 value of the prediction significantly improves as *k* increases from one to three, and afterward the trend becomes relatively steady, which suggests the terahertz spectral properties in different concentrations can be sufficiently characterized by the top three frequencies. It should be noted that these frequencies are not absorption peaks, but they are frequencies that can be used to discriminate between samples. To further visualize the discriminability of the top relevant frequencies, the concentration distributions of all instances in the top three relevant frequencies, namely 1.2, 1.1, and 0.5 THz are plotted in Fig. 6, where different concentrations are shown in different colors. It is interesting to find that the instances in the same concentration tend to cluster together and instances form different concentrations are likely separated from each other in the plots. It reveals the majority of the instances are distinguishable with respect to the top three frequencies.

The results in Fig. 6 show that measurement groupings are not monotonic in terms of the concentrations. That is the reason we proposed to use SVR with a radial basis function kernel to map the original data to a higher-dimensional space. In this way, the non-linear data in the original lower-dimensional space is very likely become linear in the higher-dimensional space. MIC technique is also demonstrated to identify non-linear relationship of the data well as shown in Fig. 6.

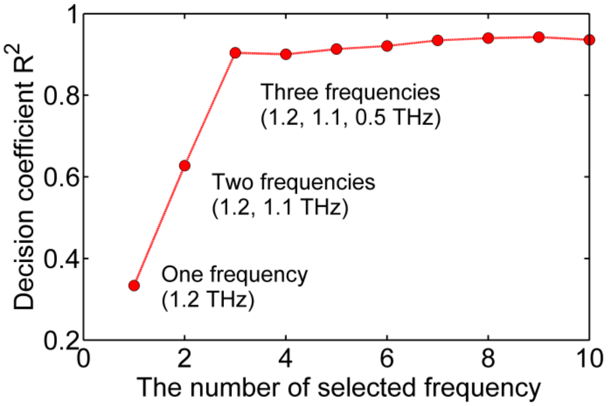

Fig. 5. The curves of R^2 with different the number of the input relevant frequencies

Table 1. The top five relevant frequencies based on MIC values

Rank	MIC value	Frequency (THz)
1	0.933470	1.2
2	0.923787	1.1
3	0.899503	0.5
4	0.890793	1.7
5	0.887096	0.4

Furthermore, the BSA was dissolved in the DI water so that no impurities exist to interrupt the BSA deposited thin-films spectra. The identified top relevant frequencies can be recognized as the feature frequencies of BSA proteins themselves in the terahertz region, which are particularly affected by the varying quantity of protein. To highlight the features of these discriminating frequencies and illustrate how the collective vibration modes respond to these frequencies into the BSA protein, the long-wavelength elastic vibration model of a spherical particle [44] is adopted, assuming BSA is a globular protein. The frequencies of spheroidal oscillations are derived as functions of the particle radius R and multipole degree l. The frequencies of spheroidal vibrations v_s are found to be the formula $v_s = v_0[2(2l + 1)(l-1)]^{1/2}$, where v_0 stands for the basic frequency of a spheroidal deformation mode of an elastic sphere, with $v_0^2 = \mu/(\rho_0 R^2)$, where μ and ρ_0 are the shear modulus and the bulk density, respectively. In this framework, the model parameters used for obtaining the theoretical calculation are given in ref [45,46]. Accordingly, the frequency $v_0 = 0.3$ THz is obtained for the lowest mode which is closer to the two discriminating frequencies at 0.5 and 0.4 THz in Table 1. Whence, for $l = 2$ and $l = 3$, the calculated frequencies are found at 1.1 and 1.8 THz which could be associated with the other three most discriminating frequencies, namely 1.1, 1.2 and 1.7 THz in the top five relevant frequencies. This result deduces that the top few discriminating frequencies identified by MIC are approximately closer to the fundamental vibration frequencies according to a spheroidal deformation mode of an elastic sphere with the varying dipole excitation order. Here, the discrepancy can be attributed to the non-spherical shape of the BSA which leads to the uncertainties in the predictions of frequencies calculation. Meanwhile, the performance of MIC can be considerably improved by increasing the new instance in the training data.

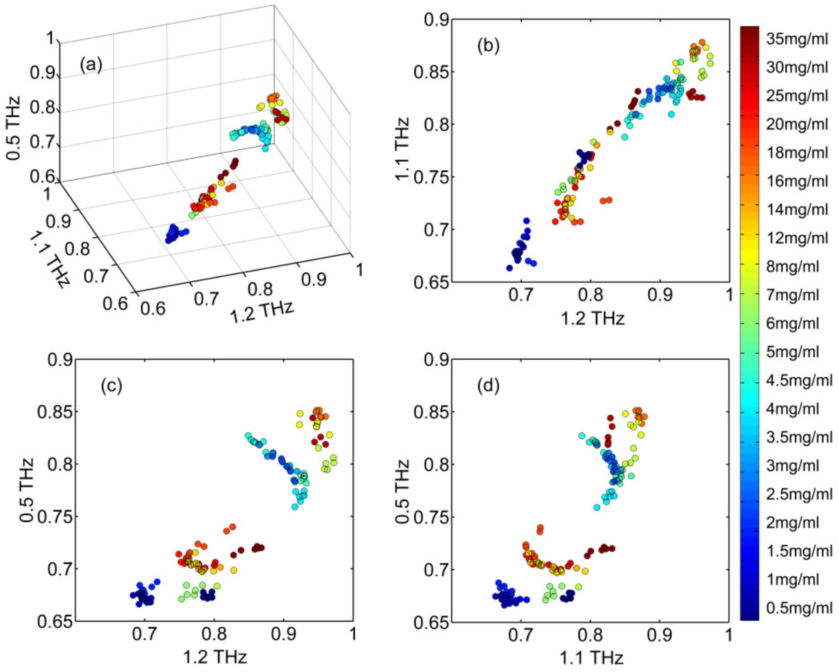

Fig. 6. The distribution of the instances in (a) the 3D-space of the top three relevant frequencies, and in (b)(c)(d) the corresponding 2D projections in the top two frequencies respectively. The units for the x, y, and z axes are normalized units represent the amplitude of the complex transmission coefficient of the samples.

4. Conclusion

In this paper, we employed terahertz time domain spectroscopy for the first time to probe BSA deposited thin-films prepared using solutions with concentrations ranging from 0.5 to 35 mg/ml. Based on the PCA denoising method, the valid data set of THz transmission coefficient spectra was input to learn the underlying model of the frequencies against the concentrations by the support vector regression method. The learned mode accurately predicts the concentrations of the unknown test samples with a coefficient of determination of $R^2 = 0.97272$. Furthermore, the maximal information coefficient is applied and three most relevant frequencies to the target concentrations are identified at 1.2, 1.1, and 0.5 THz, respectively. This means that a good prediction for BSA concentration can be achieved by using the top three relevant frequencies further proves the efficiency and practicability of the terahertz spectroscopy and machine learning methods. Additionally, these frequencies can be associated with the fundamental vibration frequencies for BSA protein according to a spheroidal deformation mode of an elastic sphere by varying dipole order. It denotes the different quantity of protein associated with a surface in a thin-film state can induce the reorientation changes accompanied by vibrational energy distribution. This result further indicates that the origin and intrinsic properties of BSA protein detected by terahertz spectroscopy can be uncovered and highlighted particularly by machine learning methods.

Our measurements and modeling highlights the unique capabilities of machine learning methods for extracting obscure characteristics from terahertz spectra for bio thin-films. Our results provide further evidence that terahertz spectroscopy in combination with machine learning methods is a sensitive analytical tool to evaluate quantity of deposited proteins in thin film systems for quality control and monitoring in the future.

Funding

National Natural Science Foundation of China (61575125); the Australian Research Council (Discovery Project, DP 150100635); the Guangdong Foundation of Outstanding Young Teachers in Higher Education Institutions (YQ2015141); Guangdong Special Support Program of Top-notch Young Professionals (2015TQ01R453); Shenzhen Scientific Research and Development Funding Program (JCYJ20150324141711587).

Disclosures

The authors declare that there are no conflicts of interest related to this article.